Taxonomía Computacional. Identificación de "Moving Groups". Aplicación a Cúmulos Abiertos

Gregorio Perichinsky[1,3] , Rosa Beatriz Orellana[2] y Ángel Luís Plastino[3].
(1) Facultad de Informática, UNLP. E-mail: gperichinsky@acm.org
(2) Facultad de Ciencias Astronómicas y Geofísicas, UNLP e *Instituto de Astrofísica de La Plata (CCT La Plata - CONICET)* E-mail: rorellan@fcaglp.fcaglp.unlp.edu.ar
(3) Facultad de Ciencias Exactas, Instituto de Física, Universidad Nacional de La Plata (CCT La Plata Laboratorio Protem – *CONICET)* E-mail: Plastino@fisica.unlp.edu.ar

CONTEXTO

- **EL PROGRAMA DE INVESTIGACIÓN CIENTÍFICA (PIC).**

  - ❖ Tiene en común un conjunto de hipótesis fundamentales, a fin de lograr un ajuste de una sucesión de teorías y resultados experimentales T1, T2, T3,…,Tn.

- En las investigaciones interesarse por algo o para avanzar en determinada dirección que trasciende el contexto del conocimiento de una disciplina, requiere:
  - ✓ Un Marco teórico.
  - ✓ Una o más teorías.
  - ✓ Homogéneas o heterogéneas.
  - ✓ Generar a través de sus hipótesis los interrogantes.
  - ✓ Mediante las conjeturas de las hipótesis formular con claridad el problema que se va a investigar.
  - ✓ Proceder a buscar su solución.

- La implicación y la generalización de los problemas y preguntas se pueden ordenar en grados: gradación.

- **ESTRUCTURA CIENTÍFICA DEL PROBLEMA**.

- La Epistemología como LA INVESTIGACIÓN EN CIENCIAS pues expone que el investigador, a través del método científico logra un conocimiento verificable, al cumplir con las etapas de la investigación científica y la etapa de contrastación de hipótesis.
- La Taxonomía Numérica para la clasificación mediante el Clustering o análisis estructural del agrupamiento de objetos y la Evidencia Taxonómica para la selección de objetos. El Taxón, las OTUs, los Invariantes y el Método SAHN.
- Los campos y herramientas emergentes de la Inteligencia Artificial, por su convergencia en la resolución del problema de clasificación automática como Aprendizaje Automático, Data Mining, Redes Neuronales y Teoría de Onditas (WAVELETS).
- Bases de Datos Relacionales Dinámicas Conceptos, Modelización, Dinámica, Generalización, Contrastación, Análisis de requerimientos, Estructuras de datos del manejador, Metaproducciones, Hiperreglas para representar las Matrices de Datos, Matrices de Similitud y Dominios en Analogía Taxonómica y medir la robustez del método con Data Mining.
- La Matemática, Estadística, Métricas (Euclídea, Manhattan y Minkowski) y Funciones de Distribución con Programación Lineal y Dinámica (Teorema de Tchebycheff).
- La Física, Espacio de Configuración, Mecánica Estadística Clásica y Cuántica (Entropía), el Electromagnetismo y Espectroscopía. Principios y Leyes.

- La Teoría de la Información, Códigos, Alfabetos, Espacio y Distancia de Hamming y el Asociador Lineal. Cantidad de Información y Entropía.
- Astronomía, Astrofísica, Astrometría, Problema de dos y muchos cuerpos, Sistema Solar, Asteroides, Estrellas y Galaxias. Grupos en movimiento.
- Ciencias de la Computación y la Informática. Ingeniería de Software y de Requerimientos.

Las Instituciones que sirvieron de sustento para este Proyecto comienza con la Universidad de Buenos Aires Facultad de Ingeniería Laboratorio de Sistemas Operativos y Bases de Datos del Departamento de Computación, donde se sentaron las bases de Bases de Datos Relacionales Dinámicas y la Taxonomía Numérica computarizada para superar el hito de los años 50´s, con la Universidad Nacional de La Plata, Facultad de Ciencias Exactas Instituto de Física Laboratorio Protem, Departamento de Informática luego Facultad de Informática y con la Facultad de Ciencias Astronómicas y Geofísicas Instituto de Astrofísica, Astrometría. Permitieron superar el hito de los 90´s donde se computarizó pero arbitrariamente y con errores, este NUEVO CRITERIO surge a partir fines de los 90´s y a partir de los 2000.
La Epistemología sirvió fundamentalmente para sentar las bases del conocimiento científico, marcar las líneas de frontera de la ciencia de la no ciencia para el PIC mostrado y contrastar tanto las Bases de Datos Relacionales Dinámicas (Alemania), como la primera constrastación del Nuevo Criterio en (UNLP Facultad de Ciencias Naturales y Museo) con el conjunto de familias en Botánica, para corroborar, la clasificación realizada en "Introducción a la Teoría y Práctica de la Taxonomía Numérica", género Bulnesia y sus Especies (Zygophyllaceae) y (Austria) con aplicaciones con Asteroides como cadáveres espaciales para la cosmogonía del Sistema Solar.
Proyectos acreditados UBACYT y PIP 6373 del CONICET.
A partir de noviembre de 2008 se continúa con el Proyecto en la Universidad Nacional de La Plata, Facultad de Ciencias Exactas Instituto de Física Laboratorio Protem con la Facultad de Informática

RESUMEN.

En este Proyecto se describen las características generales de un Nuevo Criterio para resolver el problema de la construcción de familias de objetos, mediante un programa de investigación científica (PIC). Con la Taxonomía Numérica se agrupan objetos, unidades taxonómicas operacionales en clusters (OTU´s o taxones o taxa), usando el análisis de estructura por medio de métodos numéricos. Estos clusters constituyen familias y surgen del análisis estructural, por su característica fenotípica. La Entidades formadas por dominios dinámicos de atributos, pueden cambiar de acuerdo, a los requerimientos taxonómicos: se forman familias o clusters por Clasificación de objetos, por métodos computacionales sin solapamiento entre familias, delimitación por invariantes aplicando el teorema de Tchebycheff y el máximo de la inecuación de Bienaymé-Tchebycheff, despues de normalizar el rango y con el principio de entropía máxima sobre espectros de objetos, ***original***, y aplicando principio de interferencia y superposición.
Se identifican miembros de un cúmulo abierto, con "un método espectral" agrupándose estrellas en clusters. Los caracteres utilizados para el análisis son la posición y el movimiento propio de todas las estrellas de la región del cúmulo. El método ha sido aplicado al Cúmulo NGC2516. La lista resultante de miembros concuerda muy bien con la obtenida aplicando otros métodos.

Palabras clave: Bases de Datos Relacional Dinámica, Taxonomía Numérica, Clasificación, Clusters-familias, Evidencia Taxonómica, Teoría de la Información, Espacio de Hamming, Distancia Euclídea, Espectroscopia-Principios, Teorema de Tchebycheff, Inecuación de Bienaymé-Tchebycheff, Entropía, Astrometría, Cosmogonía-Sistema Solar, Data Mining-TDIDT.

1. INTRODUCCIÓN

En este Proyecto se describen las características generales de un Nuevo Criterio para resolver el problema de la construcción de familias de objetos, mediante un programa de investigación científica (PIC). Con la Taxonomía Numérica se agrupan unidades taxonómicas operacionales en clusters (OTU´s o taxones o taxa), usando el análisis de estructura por medio de métodos numéricos. Estos clusters constituyen familias y surgen del análisis estructural, por su característica fenotípica. La Entidades formadas por dominios dinámicos de atributos, pueden cambiar de acuerdo, a los requerimientos taxonómicos: se forman familias o clusters por Clasificación de objetos. Los objetos Taxonómicos son representados mediante la aplicación de la semántica del modelo de Base de Datos Relacional Dinámica. Exhibiendo las relaciones en lo que se refiere a las calidades de similitud de los OTU´s, usando distancias Euclídeas y las técnicas de vecinos más cercanos. Así surge la evidencia taxonómica al cuantificar la similitud de cada par de OTU´s (método pair-group), obtenida de la matriz de datos básica e incorporando el concepto principal de espectro de los OTU´s, introducido originalmente, con las distancias (calculadas en base al estado de sus caracteres) y armando la matriz de similitud sobre los espectros se aplican el principio de interferencia y superposición, surgiendo el concepto de los espectros de familias, las cuales surgen, si se delimitan los grupos, a través del teorema de Tchebycheff y del máximo de la inecuación de Bienaymé-Tchebycheff, que determina Invariantes (el centroide, varianza y radio), despues de normalizar el rango y con el principio de entropía máxima. El Nuevo Criterio taxonómico es así establecido con acercamiento a la Taxonomía Computacional y presentando una explicación científica. Por otra parte se emplea con Data Mining, mediante técnicas de Machine Learning, en particular algoritmos de C4.5 de Quinlan, el grado de eficacia logrado por los algoritmos de la familia de TDIDT cuando genera modelos válidos de datos en los problemas de clasificación con la Ganancia de Información a través del Principio de la Entropía.

El **MÉTODO ESPECTRAL**, desarrollado in extenso en Perichinsky et al. (2000 y 2002), utiliza procedimientos taxonómicos para identificar los miembros de una familia a partir de los atributos taxonómicos propios de los objetos.

Se define el grado de similitud entre los objetos mediante una distancia Euclídea, que se obtiene a partir de los valores de los caracteres normalizados. La normalización de los valores de cada carácter se obtiene a partir de su valor medio y de su varianza. La distancia Euclídea normalizada (dn) permite visualizar el grado de similitud de una estrella con respecto a las demás en su "espectro característico" (Figuras 1 y 2), donde la ordenada representa la distancia dn. Los miembros del cluster se obtienen aplicando el teorema de Tchebycheff y la inecuación de Bienaymé-Tchebycheff a la distancias Euclídeas normalizadas, $d_n = \sqrt{k}.\sigma_d$, donde σ_d es la varianza de la distancia dn y k la constante de Tchebycheff que se calcula aplicando el Principio de Máxima Entropía (Perichinsky et al. 2003 y 2005). Todos los objetos miembros del cluster serán aquellas para las cuales $\sigma_d \leq \sqrt{2}\ \sigma_d$, (k = 2). Todos los objetos dudosos corresponderán a la región de incerteza para las cuales $\sqrt{2}\ \sigma_d < d_n \leq 2\ \sigma_d$, $(2 < k \leq 4)$.

En las Figuras 1 y 2, en el caso del cúmulo abierto (ver puntos 2 y 3), la línea indica el valor de d_n correspondiente a k = 2, denominado frontera. El espectro característico de una estrella (objeto) no miembro (Figura 2) no muestra valores de dn debajo de la línea de frontera, lo que indica que esta estrella (objeto) no está vinculada con alguna otra estrella (objeto) de la región. El espectro característico de una estrella (objeto) miembro (Figura 1) muestra valores de d_n por debajo de la línea de frontera, lo que indica que esta estrella (objeto) está vinculada con esas estrellas (objetos) de la región, teniendo todas ellas valores similares de los caracteres, constituyendo los miembros del cluster. Por lo tanto, el espectro característico de una estrella (objeto) miembro del cluster muestra claramente todos sus miembros.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO.

Los primeros estudios estadísticos de movimientos propios estelares para determinar la pertenencia a un cúmulo (cluster) fueron sugeridos por Vasilevskis en 1957. El modelo consiste en la suma de dos distribuciones gaussianas correspondientes una a las estrellas del cúmulo y la otra a las estrellas del campo. Posteriormente Sanders en 1971 estandariza el método aplicando el principio de máxima verosimilitud que constituye el "método paramétrico". Diferentes autores han realizado mejoras a este método con el fin de disminuir el error en la identificación de los miembros.
En 1990 Cabrera-Caño y Alfaro y en 1998 Galadí-Enríquez aplican un método denominado "método no-paramétrico" que consiste en determinar empíricamente la función de distribución de los movimientos propios.
En este trabajo que presentamos propone un nuevo método para identificar los miembros de un cúmulo abierto aplicando la Taxonomía Computacional a las posiciones y movimientos propios de las estrellas, Perichinsky et al. (2007 y 2010).
A este método lo hemos denominado "método espectral" debido a que como resultado final se obtiene un espectro que será común a todos los miembros del cúmulo.
Los cúmulos abiertos son concentraciones estelares donde sus miembros tienen características similares y la identificación de los mismos es necesaria para abordar, por ejemplo, estudios sobre la dinámica de la galaxia.
El método espectral, desarrollado in extenso en Perichinsky et al. (2000 y 2002), como se expresa antes, utiliza procedimientos taxonómicos para identificar los miembros de un cúmulo a partir de las posiciones y movimientos propios de las estrellas de la región.
Esta tarea se realiza agrupando las estrellas según el grado de similitud y afinidad en función de los valores de sus caracteres (posiciones (α y δ) y sus correspondientes movimientos propios (μ_α y μ_δ)).
Se le asigna a cada estrella un numero (i) y cada uno de sus caracteres se verá identificado con el numero de la estrella y otro numero (j) que variará de 1 a 4, según el carácter.
Se define el grado de similitud entre las estrellas mediante una distancia Euclídea, que se obtiene a partir de los valores de los caracteres normalizados. La normalización de los valores de cada carácter se obtiene a partir de su valor medio y de su varianza. La distancia Euclídea normalizada (dn) permite visualizar el grado de similitud de una estrella con respecto a las demás en su "espectro característico" (Figuras 1 y 2), donde la ordenada representa la distancia dn. Los miembros del cúmulo se obtienen aplicando el teorema de Tchebycheff y la inecuación de Bienaymé-Tchebycheff a la distancias euclídeas normalizadas.

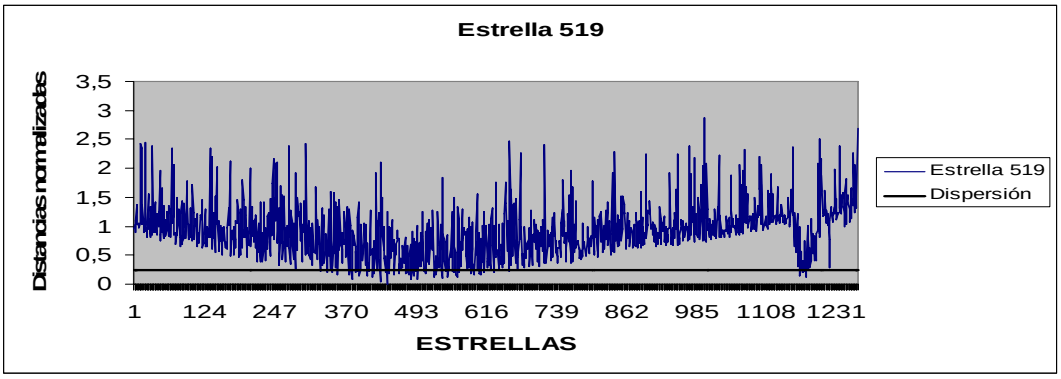

Figura 1. Espectro de una estrella miembro del cúmulo

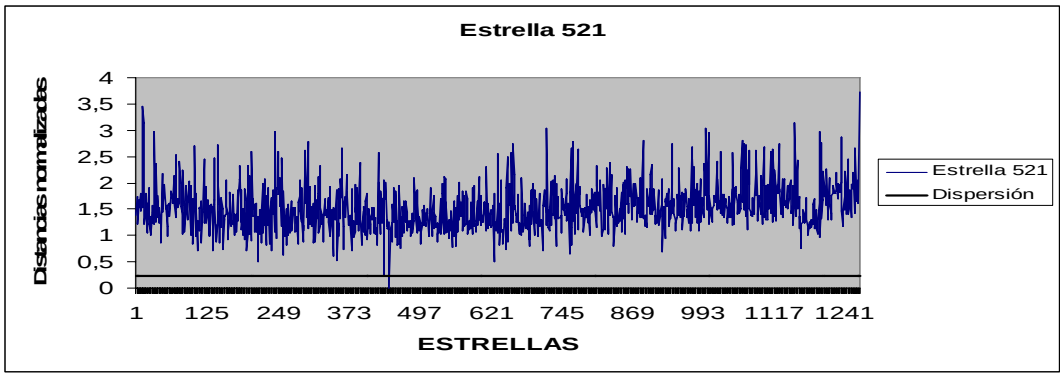

Figura 2. Espectro de una estrella no miembro del cúmulo

3. RESULTADOS OBTENIDOS/ ESPERADOS

El cúmulo abierto NGC2516 ($\alpha_{2000,0}$ = 7h 58m, $\delta_{2000,0}$ = - 60° 45') que se encuentra a unos 400 pc del sol tiene un diámetro de 35', y fue reexaminado aplicando el método espectral, se determinaron sus parámetros y se identificaron sus miembros.

Los datos de posición y movimiento propio de las 1554 estrellas de la región de NGC2516 fueron obtenidos del catálogo UCAC2 (Zacharias et al. 2004).

Una vez aplicado el método espectral se encontraron 135 miembros y los siguientes parámetros para las coordenadas y movimiento propio del cúmulo

$$\alpha = 7h\ 58m, \qquad \delta = -60º\ 45'$$

$$\mu_\alpha = -2.48 \text{ mas/año}, \qquad \mu_\delta = 9.74 \text{ mas/año}$$

El método permite también obtener el espectro característico correspondiente a cada estrella de la región. La Figura 1 muestra el espectro correspondiente a una estrella miembro del cúmulo y la Figura 2 muestra el espectro correspondiente a una estrella no miembro del cúmulo. Del análisis de todos los espectros se puede observar que existe un espectro único que caracteriza a los miembros del cúmulo por sobre las estrellas de campo, y que el mismo es similar al de la Figura 1. Es notorio que cuando una estrella es miembro su espectro atraviesa la línea correspondiente al dispersión y cuando no lo es no llega a la misma.

Cuando se compara los movimientos propios obtenidos para el cúmulo aplicando el método espectral, con los obtenidos por otros autores, se observa que nuestros resultados son coherentes con aquellos encontrados en la literatura. Ver Tabla 1.

Autores	$\mu_\alpha \cos\delta$ [mas/año]	μ_δ [mas/año]
Robinson et al.	-4.04	10.95
Baumgardt et al	- 4.08	10.98
Dias et al	- 3.2	10.1
Método Espectral	- 2.49	9.74

Tabla 1. Movimientos Propios del Cúmulo NGC2516.

Se ha presentado un nuevo método, "el método espectral", para identificar miembros de un cúmulo

utilizando las posiciones y movimientos propios de las estrellas de la región, basado en la taxonomía computacional como evolución de la numérica.
Se determinan clusters, mediante el uso de invariantes, tales como: la varianza, el radio, la densidad y el centroide, que se computan con los valores de las características de los miembros.
Se obtienen espectros característicos de las estrellas, que muestran claramente la estructura de la distribución y el agrupamiento de las mismas en clusters.
Ha sido aplicado satisfactoriamente al cúmulo NGC2516.
Es satisfactoria también la robustez del método mediante herramientas de Data Mining como C4.5 de los algoritmos de TDIDT de Quinlan con una exactitud del 3% en los casos de uso utilizados como Botánica, Medicina, Ingeniería y principalmente en Astronomía aplicación principal de problemas de Clasificación en Asteroides y Cúmulos y Galaxias.
En estos últimos casos se han hecho aplicaciones Open Clusters y se detectó las distintas formas de los mismos y un elegante poder separador de Clusters donde otros métodos en Investigación de Campos se verificaban dos Clusters con el Método Espectral se detectan tres, por lo cual seguiremos investigando en Cúmulos NGC2516, NGC3114, NGC3532 y NGC5822.
También se profundizará enana Herramienta en Ingeniería de Software para lograr un desarrollo automatizado tipo Sistema Experto para Configurar y determinar Familias de Casos de Uso.

4. RECURSOS HUMANOS

La trayectoria de los proyectos vinculados por I/D al actual, comienza con Proyectos Acreditados en UBACYT con la heurística o técnica de indagación y descubrimiento y raíz de las hipótesis de trabajo y vinculación de bases empíricas de la Taxonomía Computacional con tres tópicos: bases de datos, herramientas de inteligencia artificial o machine learning con data mining y redes neuronales y taxonomía numérica.
Con el apoyo y coordinación del Prof. Dr. Luís Ángel Plastino se desarrollaron temas de los mencionados, por problemas planteados dentro del Programa Protem del CONICET del Laboratorio del mismo nombre del Instituto de Física de la UNLP y además con la Explicación Científica de la Prof. Dra. Rosa Beatriz Orellana del Instituto de Astrofísica y del CONICET, sobre todo en tópicos de Astrometría y concluyendo en la solución de planteos de Clasificación.
Con Prof. Dr. Gregorio Perichinsky como Director de los proyectos de UBACYT y los Profs. Dr Luís Ángel Plastino y Dra. Rosa Beatriz Orellana como Apoyo Científico Externo de los mismos, han sido miembros, a su vez, más de 20 investigadores formados y en formación, miembros en los distintos proyectos. Dentro de los proyectos realizados se desarrollaron varias Tesinas y Tesis Carreras de Especialidad, Magíster y Doctorado en Ingeniería Informática y de Sistemas, Tecnología Informática Aplicada a la Educación y Ciencias Informáticas. Algunas están en curso.
Desde 2008 se finalizaron una tesis de Doctorado en Ciencias Informáticas y otra está en curso, dos de Especialidades en Tecnología Informática Aplicada a la Educación y están en curso dos de Magíster en la misma orientación y una Tesina en Ingeniería en Informática.
La estructura de los equipos de trabajo han constituido líneas de I/D en Sistemas Operativos y Bases de Datos, Inteligencia Artificial y sus herramientas, Ingeniería de Software, Tecnología Informática Aplicada a la Educación, Taxonomía Numérica y Clasificación, Teoría de la Información y Sistemas Complejos y Dinámicos.
A fines de 2008 el Prof. Dr. Gregorio Perichinsky integró el proyecto al Programa Protem del CONICET bajo la Dirección del Prof. Dr. Luís Ángel Plastino en la UNLP y mantiene contacto con los miembros del grupo de la Facultad de Ingeniería de la UBA, con tres investigadores formados y un alumno realizando la tesina de grado en Ingeniería Informática, más dos Investigadores en formación.

5. BIBLIOGRAFÍA

Baumgardt et al., 2000, A&AS, 146, 251
Cabrera-Caño y ALfaro, 1990, AJ, 63, 387
Dias et al 2006, A&A, 446, 949
Galadí-Enriquez 1998, AJ, 63, 387
Perichinsky et al. 2000, Spectra of Taxonomic Evidence in Databases.II. Application in Celestial Bodies. Asteroids families. Proceedings of XVIII International Conference on Applied Informatics. Innsbruck. Austria.
Perichinsky et al. 2003, Taxonomic Evidence Applying Algorithms of Intelligent Data Mining. Asteroids families.Proceedings of International Conference on Computer Science, Software Engineering, Information Technology, e-Business and Applications (CESITeA´03). Rio de Janeiro. Brazil.
Perichinsky et al., 2005, Electronic magazine of Systems of Information (RESI). Edition 6 - Year IV - Volume IV - Number 2. Department of Computer Science and Statistic. Federal University of Santa Catarina. Brazil.
Perichinsky et al., 2007, Taxonomic evidence of classification applying intelligent data mining. Galactic and Globular Clusters.Journal of Engineering. Tome V Fascicule 2. ISSN 1584 – 2665. University “Politechnica” Timisoara Faculty of Engineering–Hunedoara. Rumania.
Perichinsky et al., 2010, Computational Taxonomy for identification of "Moving Groups". applied to open clusters. Journal of Engineering. Tome V Fascicule 2. ISSN 1584 – 2665. University “Politechnica” Timisoara Faculty of Engineering–Hunedoara. Rumania.
Robichon et al 1999, A&A, 345, 471
Sanders,WL, 1971, A&A, 14, 226.
Vasilevskis, S et al. 1958, AJ, 63, 387
Zacharias et al. 2004, AJ, 127, 3043.